Simulating Human-to-Human Conversations for the Prediction of Conversational Quality

Thilo Michael[1], Sebastian Möller[1,2]
[1] *Quality and Usability Lab, Technische Universität Berlin*
[2] *German Research Center for Artificial Intelligence (DFKI), Berlin, Germany*
Email: [firstname.lastname]@tu-berlin.de

Abstract

Degradations in current telephone transmissions like packet-loss or delay have a multitude of repercussions on a conversation. Current instrumental estimations of transmission quality with delay and packet-loss do not take into account the type of conversation that is being measured. In this paper, we propose a new approach to predict the perceived quality of transmissions affected by delay and packet-loss by simulating the impacts these degradations have on a conversation. For this, we use user simulation techniques from the spoken dialogue community to model human conversational and turn-taking behavior in two types of scenarios, namely Short Conversation Tests and Random Number Verification Tests. We show how the impact of packet-loss may be modeled by simulating the misunderstanding of information, and the respective restructuring of the conversation to resolve that misunderstanding. We also propose a way intended and unintended interruptions can be simulated with turn-taking rules, and how the simulated agents might resolve the conflicts that arise from these interruptions.

We then outline possible quality measures that may be derived from such a system, including a mean opinion score for degraded conversations.

Introduction

The estimation of quality of speech over a telephone network has long been a topic in the research community. The goal is to create speech technology that has an optimal quality while using the least amount of resources. Besides subjective evaluation methods like listening tests, speaking tests and conversational tests, that allow quantification of the perceived quality, parametric estimation of telephone quality is an important factor, especially for the planning phase of telephone systems and their optimization.

The introduction of packet-based telephone systems combined with significant amount of signal processing in the terminal device has made delay and packet-loss a focus of research. While delay itself cannot be noticed in listening tests, they have a significant impact on the turn-taking modalities of a conversation and thus influence the perceived quality. Also packet-loss is shown to have a context-specific influence. Depending on which part of an utterance is affected by packet loss, the interlocutor may have to communicate about that misunderstanding, which impacts the structure of the conversation.

In contrast to the evaluation of speech quality, the conversational quality depends on interactive factors. This includes not only the types and severity of degradations, but also the interactivity of the conversation and even the personality of the interlocutors is an important factor of the quality rating. Especially turn-taking and meta communication influence the perceived quality of a degraded conversation, which parametric quality predictions cannot accurately account for.

Recent work in the areas of user simulation and turn-taking prediction in spoken dialogue systems have made an approach feasible, where task-oriented conversations can be modeled on a semantic level. For this, the utterances of each interlocutor is abstracted as dialogue acts that represent the intent of the speaker and concepts or slots, that represent information that is being exchanged with that utterance. In the field of incremental (that is *real-time*) spoken dialogue systems, turn-taking and end-of-turn prediction is proposed to precisely time the utterances of the system so that a natural conversation flow can be created.

Using these approaches a simulation of a human-to-human conversation may be modeled that represents the dialogue on signal level, on text level and also on abstract dialogue act level. Such a simulated dialogue would not only conform to the timely characteristics of a conversation but it would also model the semantic communications. With the framework of a simulated conversation degradations like packet-loss and delay may be introduced. The turn-taking and dialogue management units would then be able to react to those degradations accordingly. In the structure of the conversation and the meta communication caused by the degradations, such a simulation would show similar characteristics as empirical data under same conditions. This way a prediction of the overall perceived quality of the simulated conversation may be achieved.

In this paper we outline a simulation of human-to-human conversations and how specific characteristics on signal level as well as on the semantic level have to be modeled to reproduce behavior seen in degraded conversations. We show how delay may affect the simulation framework and how human behavior in regards to delayed conversations may be modeled. We show how the framework may be modeled to react to mishearing due to packet-loss and to resolve conflicts that might arise. We outline a way how the results of such a simulation may be used to predict the conversational quality.

Related Work

In the following we will outline the related in both the measurement of conversational quality and tests and in the area of user simulation and turn-taking prediction in spoken dialogue systems.

Conversational Quality

Subjective evaluation of telephone quality [1] and especially the conversational quality [2] has been a research focus for a long time. To assess the quality - especially in regards to the flow of the conversation - conversation tests like the Short Conversation Test (SCT) [3], the interactive Short Conversation Test (iSCT) [4] and the Random Number Verification (RNV) [5] are used. To better predict and model conversational quality, the conversation can be separated into three phases: the *listening* phase, the *speaking* phase and the *interaction* phase [6]. These phases can be evaluated separately and can be used to better model the conversational quality [7, 8].

While degradations like packet-loss influence the shape of the speech signal and thus are detectable by tests that require participants only to listen, effects of delay are more complicated to assess [9, 18]. In [10], Hammer showed that the type of conversation - specifically the conversational interactivity - has a significant influence on the impact of delay. When a conversation has a higher conversational interactivity, delay has more impact during turn-taking and thus the perceived quality degrades more. Parametric conversational analysis defines features like double talk, mutual silence, speaker alternation rate and interruption rate, that are indicators for the conversational quality.

Packet-loss on the other hand has a detectable influence on the signal itself. However, in the context of conversational quality it has diverse repercussions [8, 11]. Packet-loss in a highly interactive conversation has a higher impact on conversational quality than in a conversation with lower interactiveness, because every mishearing of important parts of an utterance yields additional meta communication to repair that misunderstanding.

Simulations of Conversations

Simulating conversational speech is long used for the assessment of speech quality [12]. However, these simulations only include the on-off temporal characteristics of human conversational speech, where the length of the talk-spurts and the pause and double talk rate match the parameters of human conversations.

In the area of spoken dialogue systems, simulations of a conversation are used as a way to generate dialogues for evaluation or training of those systems [13]. However, user simulations only model interactions of a machine (i.e. a dialogue system) with a human. They are often based on dialogue acts and concepts, an abstract semantic representation of the intents and content of the dialogue. Because those simulations are often realized on a dialogue act basis, the user models are evaluated on the semantic level [14, 15]. Many metrics have been used to capture the similarities between simulated and real dialogues on global, dialogue-wide level as well as on turn-based level [16].

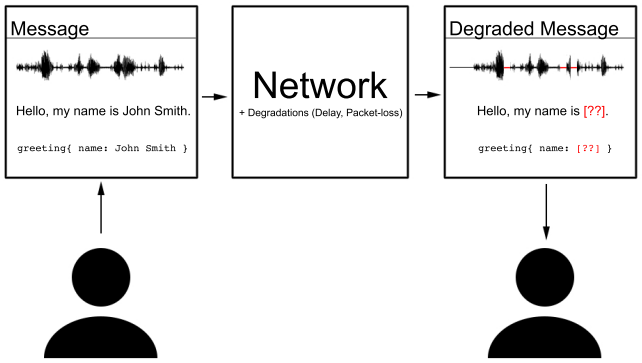

Figure 1: A visualization of the two simulated agents and the network. The agent on the left sends out a message containing the signal (top), the transcribed text (middle) and the intent and concepts (bottom) of the utterance. The network degrades the signal with packet-loss and delay and the effects are modeled in the transcription and the dialogue act. The receiving agent may then use dialogue strategies to act accordingly (e.g. ask for the name again).

Also the simulation of conversations is used to simulate turn-taking behavior in spoken dialogue systems [17]. While these simulations are based on a more formal description of turn-taking than described in the ITU-T recommendation P.59, they do not consider the content of the conversation as the exchanged utterances do not convey any information.

Human-to-Human Simulation

The human-to-human dialogue simulation proposed here combines the approaches of simulation with user models and simulation of turn-taking behavior. In a simulated environment two agents exchange multi-layered acts through a network. These acts are not only represented by the dialogue acts and concepts, but also contain a transcription and the speech signal itself (as shown in Figure 1). As a basis conversations with different conversational interactivity may be used, for example the Short Conversation Test and the Random Number Verification test. These types of conversational scenarios are well tested and provide for a goal-oriented semi-structured task that fits into the types of conversations simulated in user modeling.

In the semantic part of the simulation, the two agents use dialogue management to decide which act should be spoken next. This decision can be based either on the dialogue act that is transmitted, or it can be adapted depending on the audio and the text representation of the utterance. With end-of-turn and turn-taking prediction models, the agents goal is to decide when to take over the turn and when to backchannel. These turn-taking predic-

tions have to include the rule-based turn-taking as well as stochastic approaches. Together with the information about degradations and the content of the utterance, the agents need to use different turn-taking strategies. For example, taking a turn early is dependent on the content of the utterance as well as on the length (a lengthy utterance is more likely to be interrupted than a short one) [9].

The network is streaming the packets from one agent to the other. During transmission, the packet forwarding may be delayed and packets may be withheld to simulate packet-loss. In a first version of the simulation model, the networking module adds the information about what kind of degradation and where a degradation occurs in the meta data of the message. In later versions this simulated network may then be replaced by a real network and effects on speech understanding and turn-taking could be modeled inside the agents.

Delay

The agents in the simulation framework simulate turn-taking by estimating the end of turn of the interlocutor and interrupting them based on stochastic models. During the introduction of delay through the network, the agents would continue taking turns with the same model. This results in unintended interruptions that may be similar to real delayed conversations at low delay levels.

During higher delay levels, people tend to notice the delay and change their turn-taking behavior [18]. In the simulation, the two agents would need to be able to detect an increased amount of turn-taking as well. The agent could monitor the number and length of interruptions in the conversation, taking into account the varying degrees of interactivity throughout the conversation. If a certain threshold interruption length or repetition is reached, the agents could then adapt their turn-taking behavior to include an estimation of the delay.

Long delays may also alter the content of the dialogue. If an interlocutor is not answering a question (for example because the question takes too long to transmit) an agent may request a confirmation that the interlocutor is still listening. The behavior itself is not unique to conversations with transmission delay, but it increases with delay and thus the number of these instances may be a good indicator for conversational quality.

Packet-loss

To simulate human-like behavior during conversations degraded with packet-loss, the agents need to be able to detect bits of information that were impacted from the packet-loss and to resolve those misunderstandings with meta communication.

For the detection of disturbed transmission of relevant bits of information, the agents may use the meta information about the types and locations of the degradations provided by the simulated network. In further versions, the agents provide their own means of detecting and locating packet-loss and delay and are able to estimate which parts of the dialogue act and concepts was transmitted in a valid way.

To properly react to misunderstanding, the agents need to resolve them by posing questions about the last utterance of the interlocutor. For this, different contexts have to be handled by the agents - even nested context switches (e.g. when there appear misunderstandings while the interlocutors trying to resolve another mishearing). After resolving the problems, the agent need to be able to continue the conversations at the point where they left it and continue working towards the given goal.

Quality Estimation

The requirements for a quality estimation are that the results of the simulation are comparable to real-world conversations. That means for a simulation with the same networking parameters, conversation scenario and interactiveness as the real-world conversation, similar characteristics should be achieved in the conversation analysis (double talk, mutual silence, speaker alternation rate, etc.) and on the semantic level (type of dialogue acts, similarity of conversation structure).

Additionally to the measures of the conversational analysis that have been correlated with the perceived quality before [18, 9], the semantics of the conversation could be used to estimate the quality of the conversation. Because the conversation is also simulated on the semantic level through dialogue acts, the temporal structure of these acts and also the occurrences of certain types of acts can be measured. For example, the frequency of the dialogue act *misunderstanding* could be an indicator of a degraded conversation.

Different from any previous approach to instrumental prediction of conversational quality, the simulation would enable that every simulated interlocutor of every simulated conversation may generate an *artificial quality rating* that can be averaged to create a MOS value for a specific setting of the conversation. Additionally, agents could include different strategies on how to handle the problems that arise from the degradations into the quality rating to match real-world participants.

Conclusion & Outlook

In this paper we outlined a new approach to the assessment of conversational quality, namely the simulation of a conversation. We showed how such a simulation system can be implemented and what requirements the system has to fulfill to be able to appropriately simulate a goal-oriented conversation as well as problems that arise from degradations like delay and packet-loss. We describe how turn-taking may be modeled in a way that considers the content of the utterance and also how packet-loss may affect the information transmitted by an interlocutor and

how this can create misunderstandings that the simulated agents have to resolve.

In the future we will validate the implementation of such a simulation framework for agents modeled without delay and investigate how the introduction of delay will impact the structure of the simulated conversation. We will model how a simulated agent detects misunderstandings and which steps an agent takes to resolve them.

Acknowledgements

This work was supported by the German Research Foundation DFG (grant number MO 1038/23-1).

References

[1] ITU-T Recommendation P.800, Methods for subjective determination of transmission quality. Geneva, Switzerland: International Telecommunication Union, Aug. 1996.

[2] ITU-T Recommendation P.805, Subjective Evaluation of Conversational Quality. Geneva: International Telecommunication Union, 2007.

[3] S. Möller, Assessment and prediction of speech quality in telecommunications. Kluwer Academic Publishers, 2000.

[4] A. Raake, Speech quality of VoIP: assessment and prediction. John Wiley & Sons, 2007.

[5] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications", IEEE Journal on selected Areas in Communications, vol. 9, no. 4, pp. 586–593, 1991.

[6] M. Gueguin, R. Le Bouquin-Jeannes, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," EURASIP Journal on Advances in Signal Processing, vol. 2008, p. 93, 2008.

[7] S. Möller, F. Köster, and B. Weiss, "Modelling speech service quality: From conversational phases to communication quality and service quality", in Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on. IEEE, 2017, pp. 1–3.

[8] F. Köster, Multidimensional Analysis of Conversational Telephone Speech. Springer, 2017.

[9] F. Hammer, P. Reichl, and A. Raake, "The well-tempered conversation: interactivity, delay and perceptual VoIP quality", in IEEE International Conference on Communications, vol. 1. Institute of Electrical and Electronics Engineers (IEEE), 2005, pp. 244–249.

[10] F. Hammer, Quality Aspects of Packet-Based Interactive Speech Communication. Forschungszentrum Telekommunikation Wien, 2006.

[11] A. Takahashi, A. Kurashima, and H. Yoshino, "Objective assessment methodology for estimating conversational quality in VoIP", in IEEE Transactions on Audio, Speech, and Language Processing, 2006

[12] ITU-T Recommendation P.59, Artificial Conversational Speech. Geneva, Switzerland: International Telecommunication Union, 1993.

[13] W. Eckert, E. Levin, and R. Pieraccini, "User modeling for spoken dialogue system evaluation", in Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on. IEEE, 1997, pp. 80–87.

[14] K.-P. Engelbrecht, Estimating Spoken Dialog System Quality with User Models. Springer Berlin Heidelberg, 2013.

[15] S. Hillmann, Simulation-Based Usability Evaluation of Spoken and Multimodal Dialogue Systems. Springer, 2017.

[16] O.Pietquin and H.Hastie, "A survey on metrics for the evaluation of user simulations", The knowledge engineering review, vol. 28, no. 1, pp. 59–73, 2013.

[17] T. Baumann, "Simulating Spoken Dialogue With A Focus on Realistic Turn-Taking", 13th ESSLLI Student Session, pp. 17–25, 2008.

[18] S. Egger, R. Schatz, and S. Scherer, "It takes two to tango - assessing the impact of delay on conversational interactivity on perceived speech quality", in Eleventh Annual Conference of the International Speech Communication Association. ISCA, 2010, pp. 1321–1324.